



# **LSAView: A Tool for Visual Exploration of Latent Semantic Modeling**

Patricia Crossno, Daniel Dunlavy, Timothy Shead

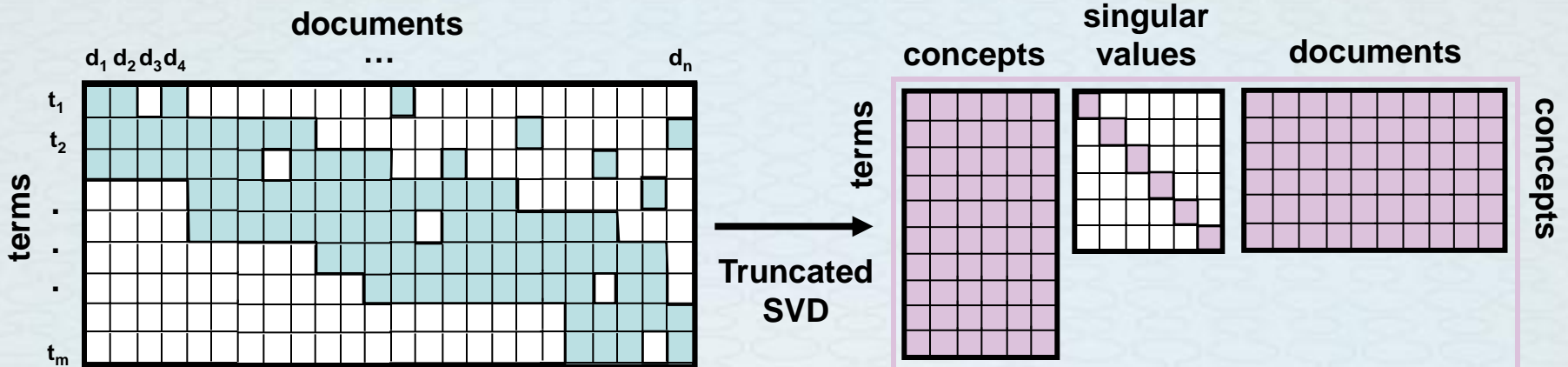
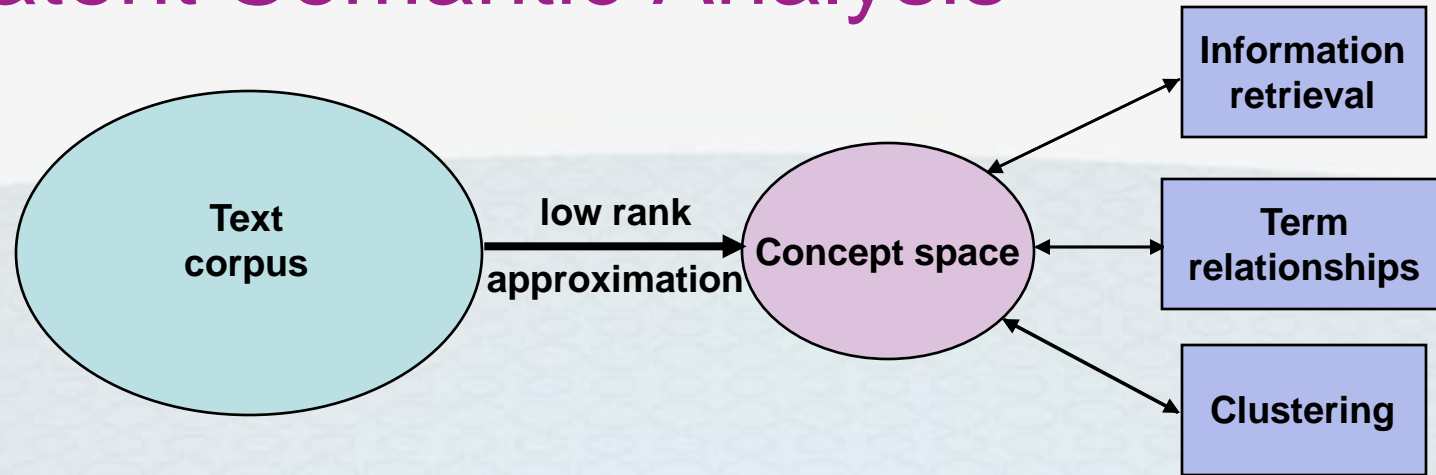
Sandia National Laboratories

# Overview

- Latent Semantic Analysis
- Motivation
- Analysis of Algorithmic Choices
- LSAView
- Case Studies
  - Rank Selection
  - Singular Value Scaling
- Conclusions



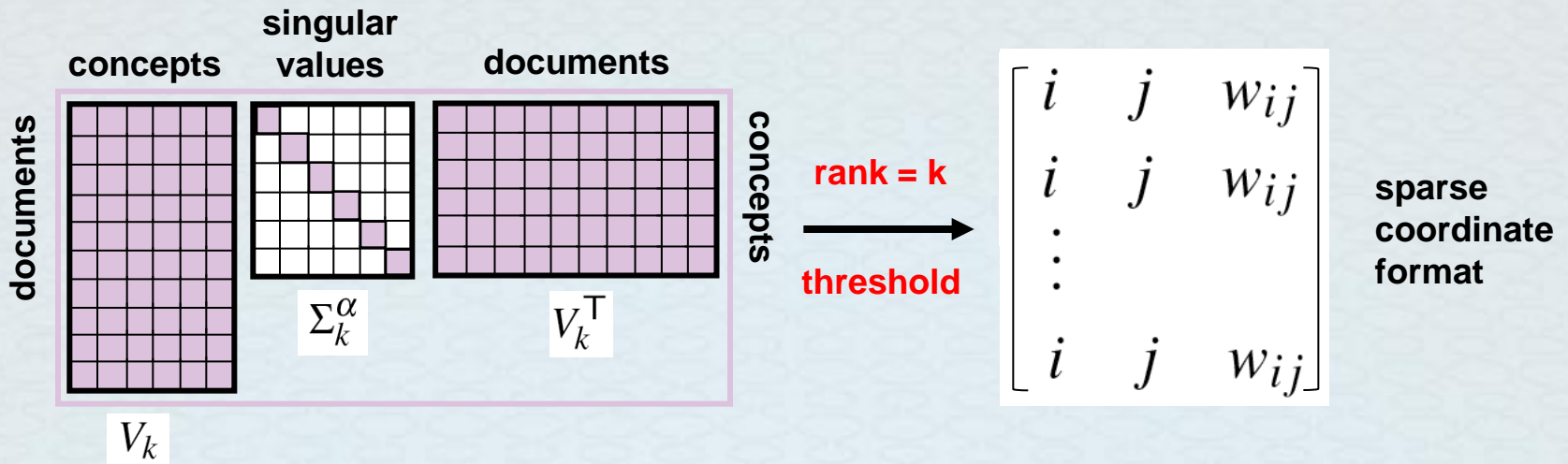
# Latent Semantic Analysis



$$A \approx U_k \Sigma_k V_k^T$$

# Document Similarity Graphs

Document similarity matrix =  $V_k \Sigma_k^\alpha V_k^\top$



Use Cosine Similarities

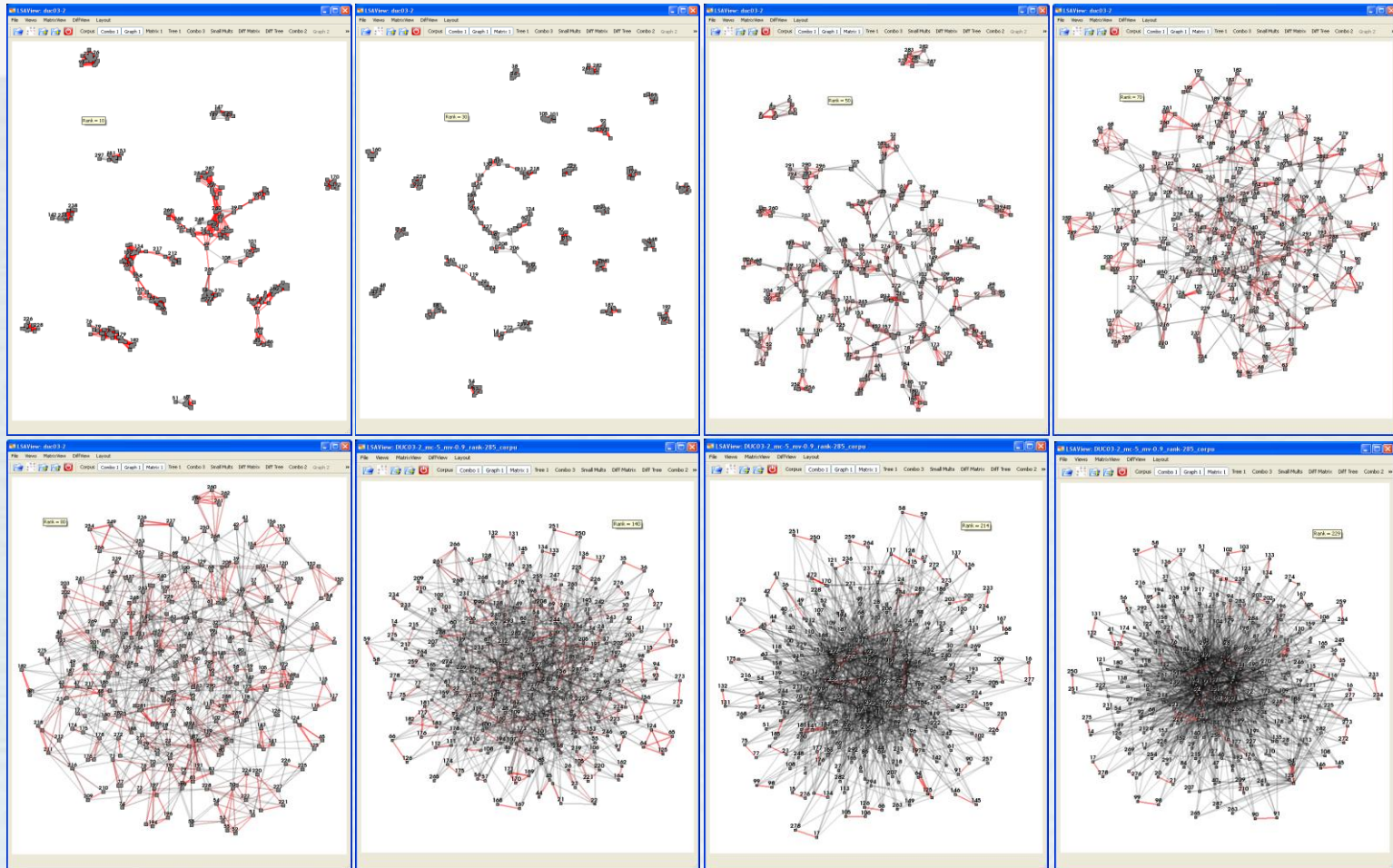
$$e_{ij}(k) = \frac{\langle v_k^i \Sigma_k, v_k^j \Sigma_k \rangle}{\|v_k^i \Sigma_k\|_2 \|v_k^j \Sigma_k\|_2}$$

Document similarity graph

- Each document is a vertex
- Each row defines an edge

# Motivation:

## Algorithmic Parameter Choices Impact Models



Which rank to use?



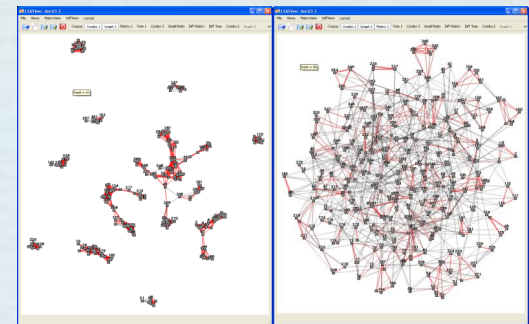
# Analysis of Algorithmic Choices

Focus on impacts from:

- Rank (number of concepts)
  - Find sweet spot between extremes
- Similarity computation
  - Singular value scaling

How to visualize model impacts?

- Conceptual groupings
  - Document layout
  - Changes in link strength between documents
- Significance of changes in edge weights
  - Large changes not necessarily significant
  - Statistical inference tests

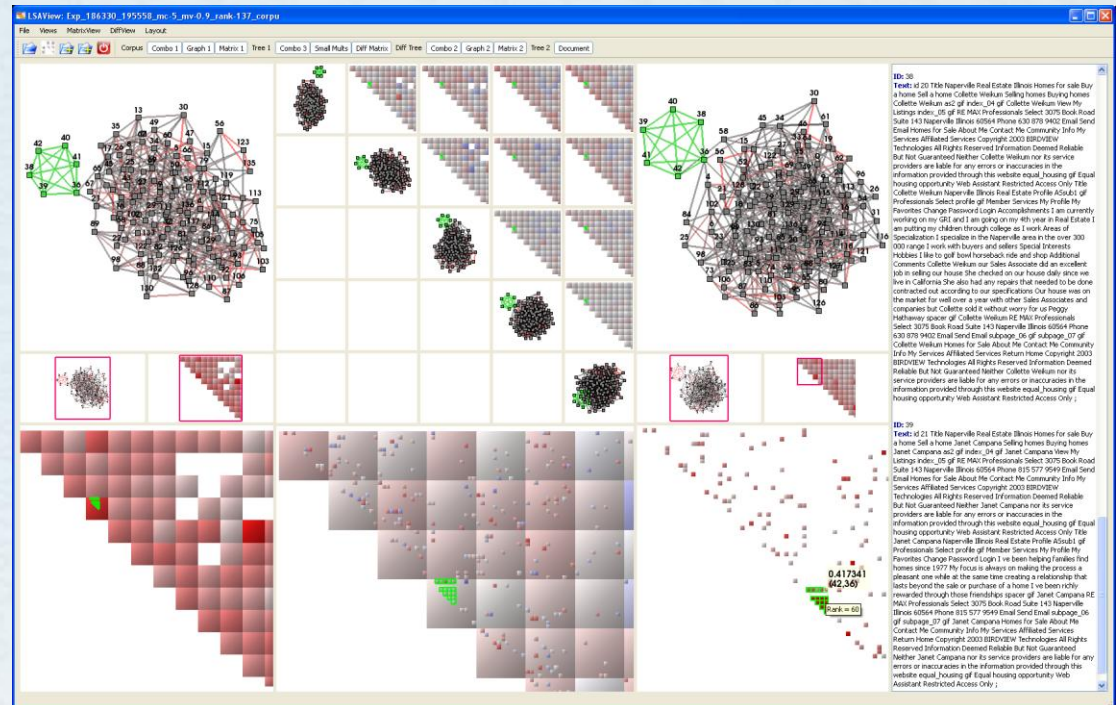


# LSAView

- Compares models
- Explores impacts of parameter choices
- Uses statistical inference to highlight model differences
- Built using open source VTK/Titan Informatics Toolkit

- Views

- Graph
- Matrix
- You Are Here
- Small Multiples
- Document



# Rank Selection Case Study

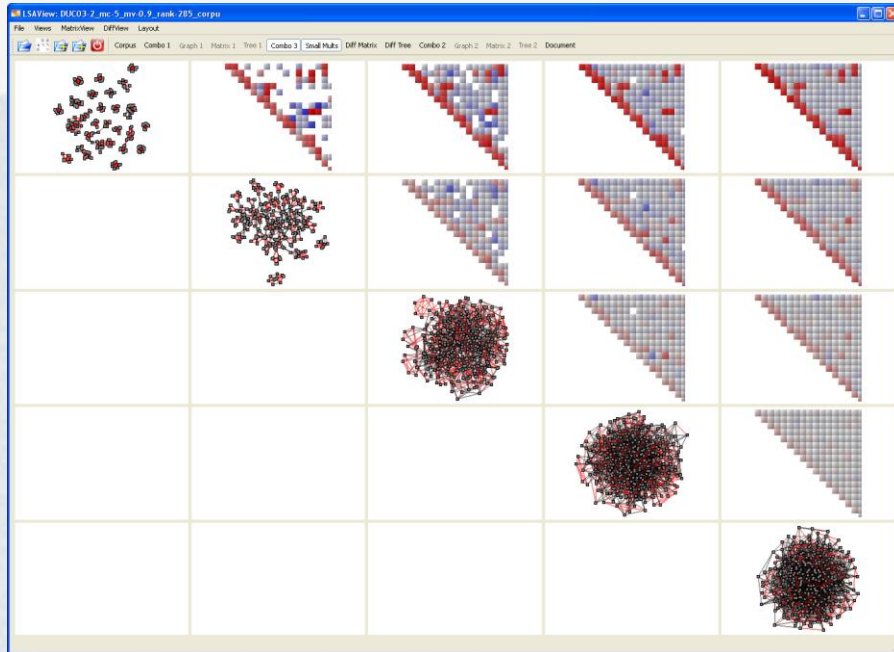
- DUC data
  - 2003 Document Understanding Conference (DUC)
  - 298 newswire documents for summarization evaluation
  - Documents in 30 clusters
  - ~10 documents per cluster on a particular topic or event
  - <http://www-nlpir.nist.gov/projects/duc/data.html>
- Rank = k (SVD truncation)

$$A \approx U_k \Sigma_k V_k^T$$
- Iterative Approach
  - Identify range of potential ranks – *Small Multiples View*
  - Compare ranks – *Graph, Matrix, and Data Table Views*
  - Validate rank – *Document View*

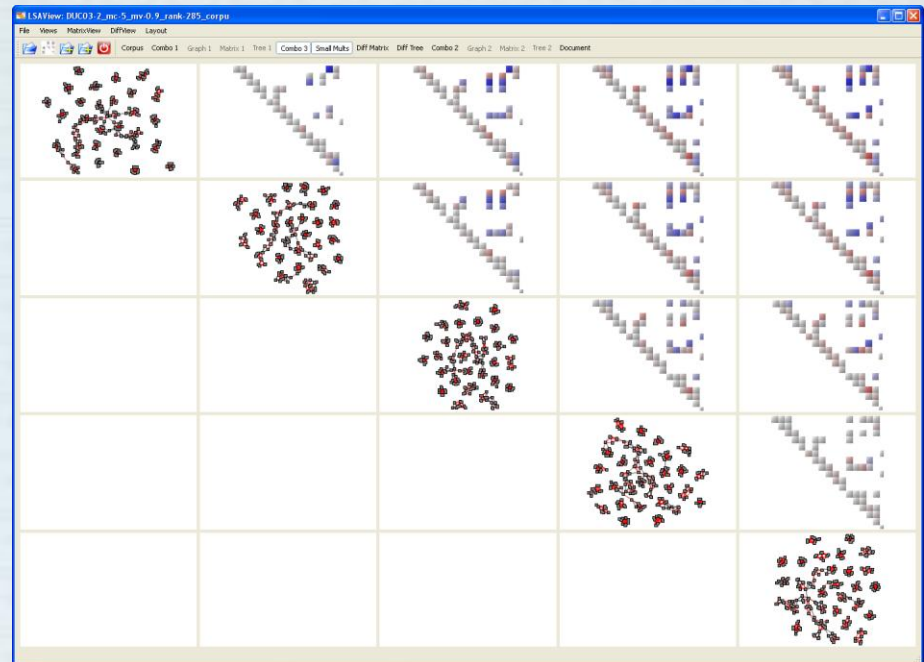




# Small Multiples: Narrow Range of Ranks



Ranks  $k = 20, 50, 80, 11, 140$

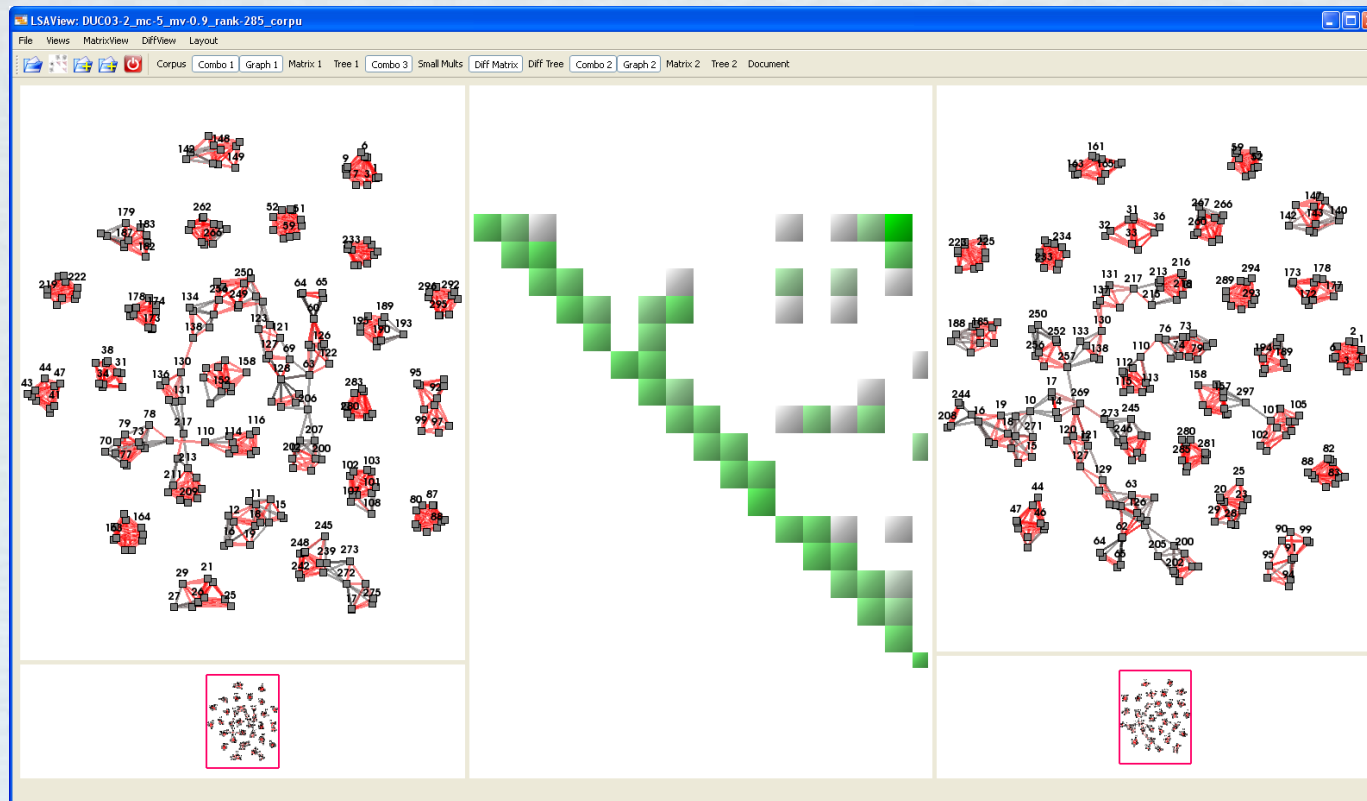


Ranks  $k = 28, 29, 30, 31, 32$

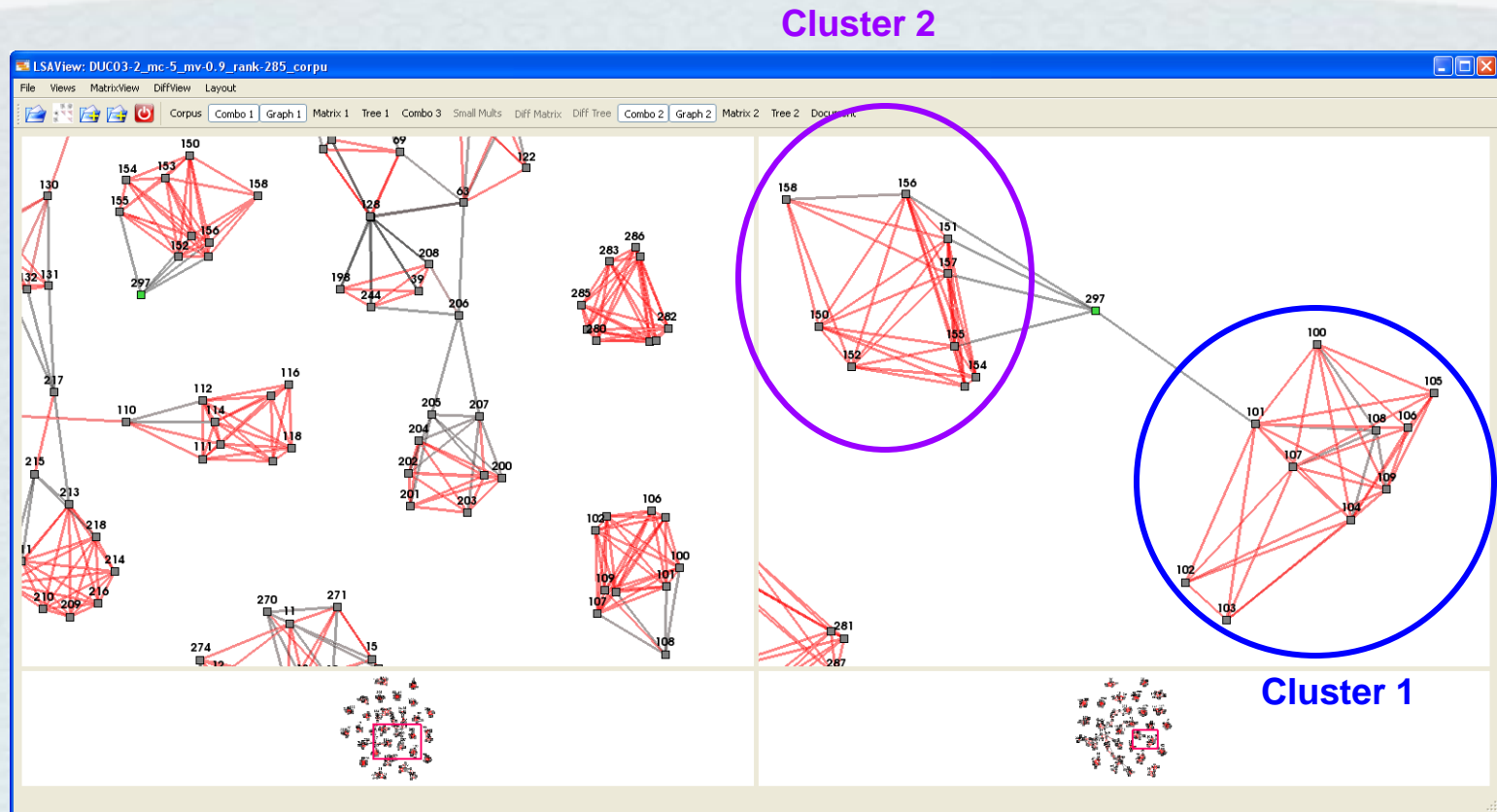
# Two-sample $t$ Statistics

$$t_{ij}^{(2)} = \frac{\bar{e}_{ij}(k_1, \alpha, n_1) - \bar{e}_{ij}(k_2, \alpha, n_2)}{\sqrt{\frac{[s_{ij}(k_1, \alpha, n_1)]^2}{n_1} + \frac{[s_{ij}(k_2, \alpha, n_2)]^2}{n_2}}}$$

- Identify anomalous edge weights between 2 graphs
- Most significant differences in bright green



# Anomalous Links to Document 297



# Manual Inspection

- Document 297 – Chinese policy on separatists
- Cluster 1 topic – trial of 3 Chinese separatists
- Cluster 2 topic – Russian policy on Chechnyan separatists
- Policy theme best match for 297, conclude Rank 30 best

The screenshot displays the LSAView application window. The title bar reads "LSAView: DUC03-2\_mc-5\_mv-0.9\_rank-285\_corpu". The menu bar includes "File", "Views", "MatrixView", "DiffView", and "Layout". The toolbar contains icons for file operations and a "Corpus" dropdown menu. The main interface is divided into two panes. The left pane shows a network graph with nodes labeled with document IDs (130, 131, 154, 155, 156, 157, 158, 198, 297) and edges representing relationships. Nodes 154, 155, 156, 157, and 297 are highlighted with green outlines, while others are in red. A small cluster of nodes is visible at the bottom of the graph. The right pane displays the text of document 297, which is a news story from 1998 about Chinese policy on separatists. The text includes the document ID, a title, a byline, and several paragraphs of news text.

LSAView: DUC03-2\_mc-5\_mv-0.9\_rank-285\_corpu

File Views MatrixView DiffView Layout

Corpus Combo 1 Graph 1 Matrix 1 Tree 1 Combo 3 Small Mults Diff Matrix Diff Tree Combo 2 Graph 2 Matrix 2 Tree 2 Document

ID: 297  
Text: NYT19981218.0250 NEWS STORY 1998-12-18 18:37 A7526 &Cx1f; taf-z u i &Cx13; &Cx11; BC-CHINA-JIANG-NYT &LR; 12-18 0890 BC-CHINA-JIANG-NYT NO OPPOSITION ALLOWED, CHINA LEADER SAYS (JT) By ELISABETH ROSENTHAL c.1998 N.Y. Times News Service

BEIJING \_ In a pointed speech delivered to Communist Party members in the Great Hall of the People, President Jiang Zemin made clear Friday that China's economic reforms were not a prelude to Western-style multiparty democracy and that those who challenged the party's supremacy would be quashed.

Officially, the speech was intended to commemorate the 20th anniversary of the Communist Party meeting that marked the beginning of the reforms. But it also provided a vigorous, if indirect, defense of the government's recent decision to try three of the country's most prominent democracy activists on criminal charges for their efforts to organize an opposition political party.

The trials of Wang Youcai and Qin Youming opened on Thursday. They are accused of inciting subversion of the state, charges that carry a minimum prison sentence of five years on conviction, although neither trial has come to a verdict.

And Friday the family of China's most senior dissident, Xu Wenli, another organizer of the opposition party, was told that he would stand trial on Monday, facing an even more serious charge of subverting state power, which could bring a sentence of life imprisonment.

Xu has been detained since Nov. 30, after the police raided his house and confiscated documents about the new opposition group, the China Democratic Party.

While vowing to move ahead with economic reforms, Jiang's speech was more conservative politically than others he has made this year. Again and again he invoked Communist Party dogma, using phrases rarely uttered in the last year, like the "four cardinal principles." Those state that the leadership of the Communist Party, adherence to the socialist path and the upholding of Mao Zedong Thought and of Marxism-Leninism must serve as the foundation for the country.

During President Clinton's state visit in the summer, the world saw President Jiang, agile international political leader, in action. Friday he was once again General Secretary Jiang, Communist Party stalwart.

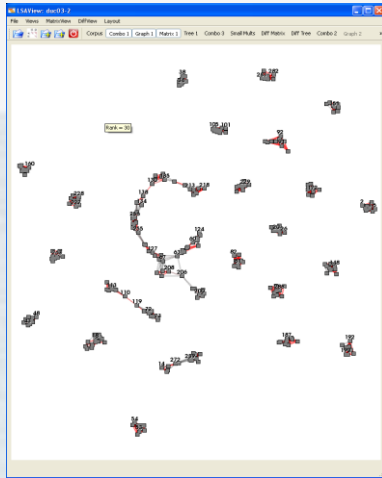
China's current political system "must not be shaken, weakened or discarded at any time," Jiang said, adding, "The Western mode of political systems must never be copied."

Later he said, "We must be vigilant against infiltration, subversive activities and separatist activities of international and domestic hostile forces." China must "fight against factors disrupting social stability and nip them in the bud," he said.

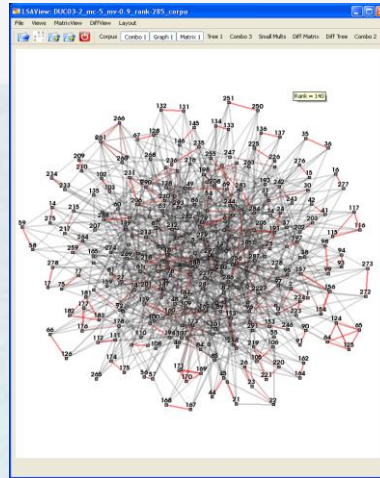
Jiang's remarks on these topics echoed points made two weeks ago by Li Peng, the former prime minister who now heads the National People's Congress, in a rare interview with the German business newspaper Handelsblatt.



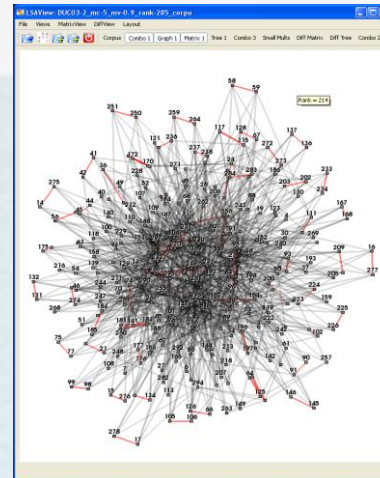
# Comparison to Automated Methods



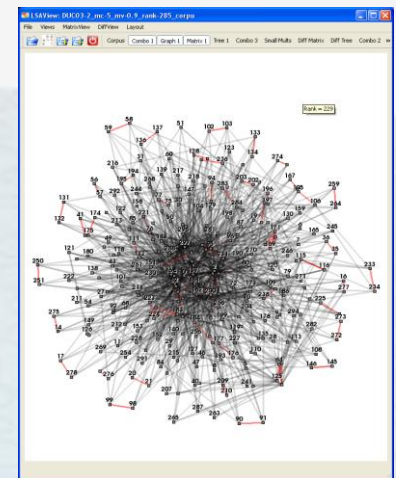
**LSAView**  
**Rank 30**  
**Variance 40.59**



**Leave-1-Out**  
**Cross Validation**  
**Rank 140**  
**Variance 80.72**



**95% Variance**  
**Rank 214**  
**Variance 95.12**



**20-group (fold)**  
**Cross Validation**  
**Rank 229**  
**Variance 97.27**

- Automated rank selection methods select ranks
  - Robust to noise
  - Accounting for variance in data
- LSAView selects on impact to text analysis tasks



# Singular Value Scaling Case Study

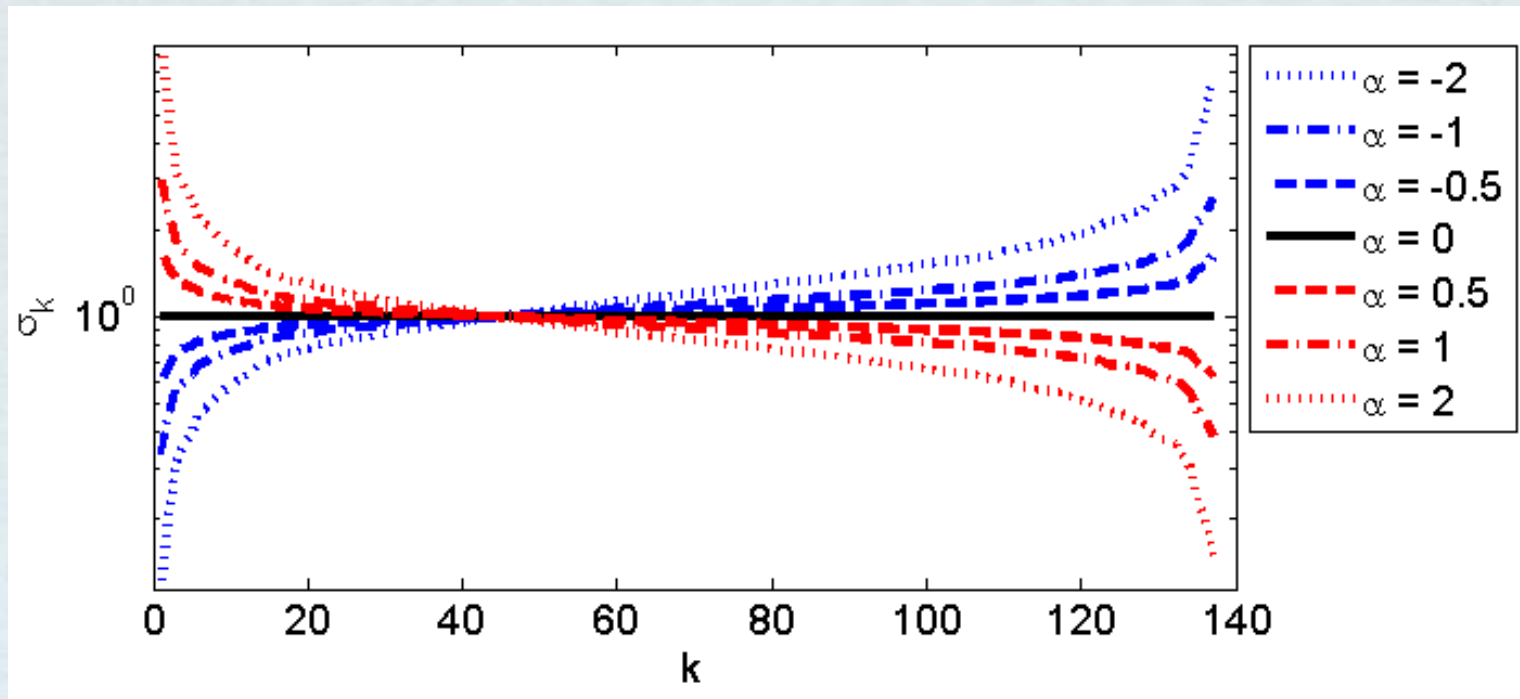
- TechTC data
  - Subset of TechTC-100 Test Collection
  - 150 html documents partitioned into 2 clusters
  - <http://techtc.cs.technion.ac.il/techtc100/techtc100.html>
- Singular Value Scaling =  $\alpha$

$$e_{ij}(k, \alpha) = \frac{\langle v_k^i \Sigma_k^{\alpha/2}, v_k^j \Sigma_k^{\alpha/2} \rangle}{\|v_k^i \Sigma_k^{\alpha/2}\|_2 \|v_k^j \Sigma_k^{\alpha/2}\|_2}$$

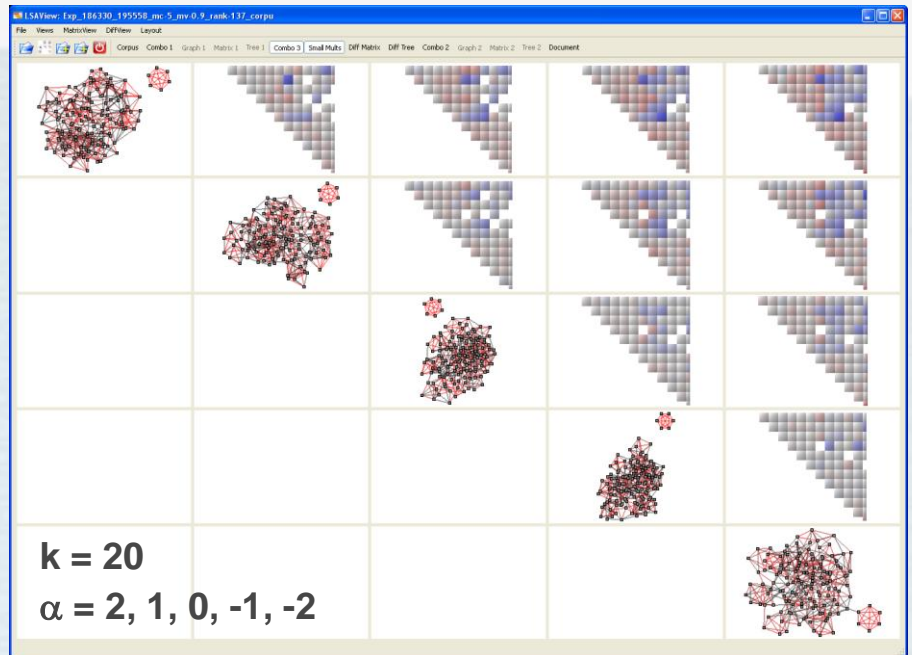
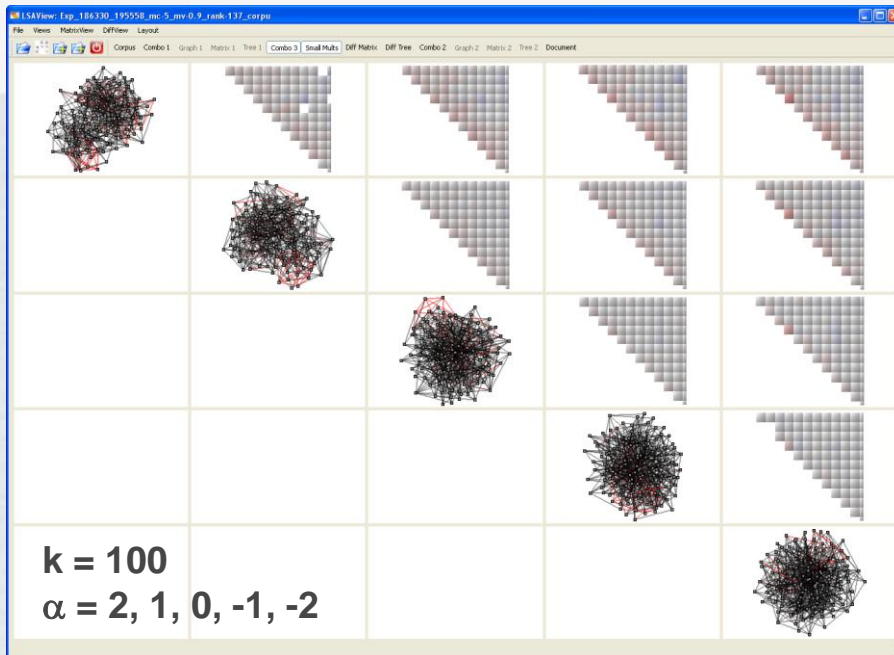
- Complicated by rank selection
- Inspect scaled singular values for  $\alpha$  vs.  $k$

# Inspect Singular Values Scaled by $\alpha$

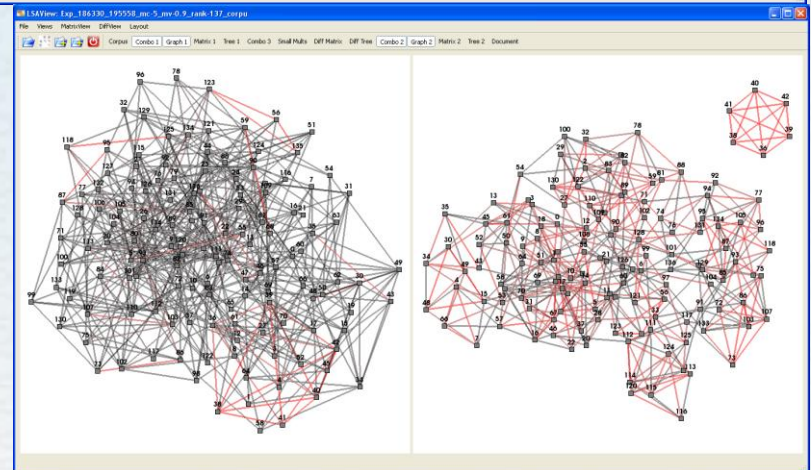
- Original singular values correspond to  $\alpha = 2$
- For all  $\alpha$ , values trend toward 0 for  $k < 45$
- For  $k > 45$ , inverted scalings amplify noise



# Small Multiples $k > 45$ vs $k < 45$

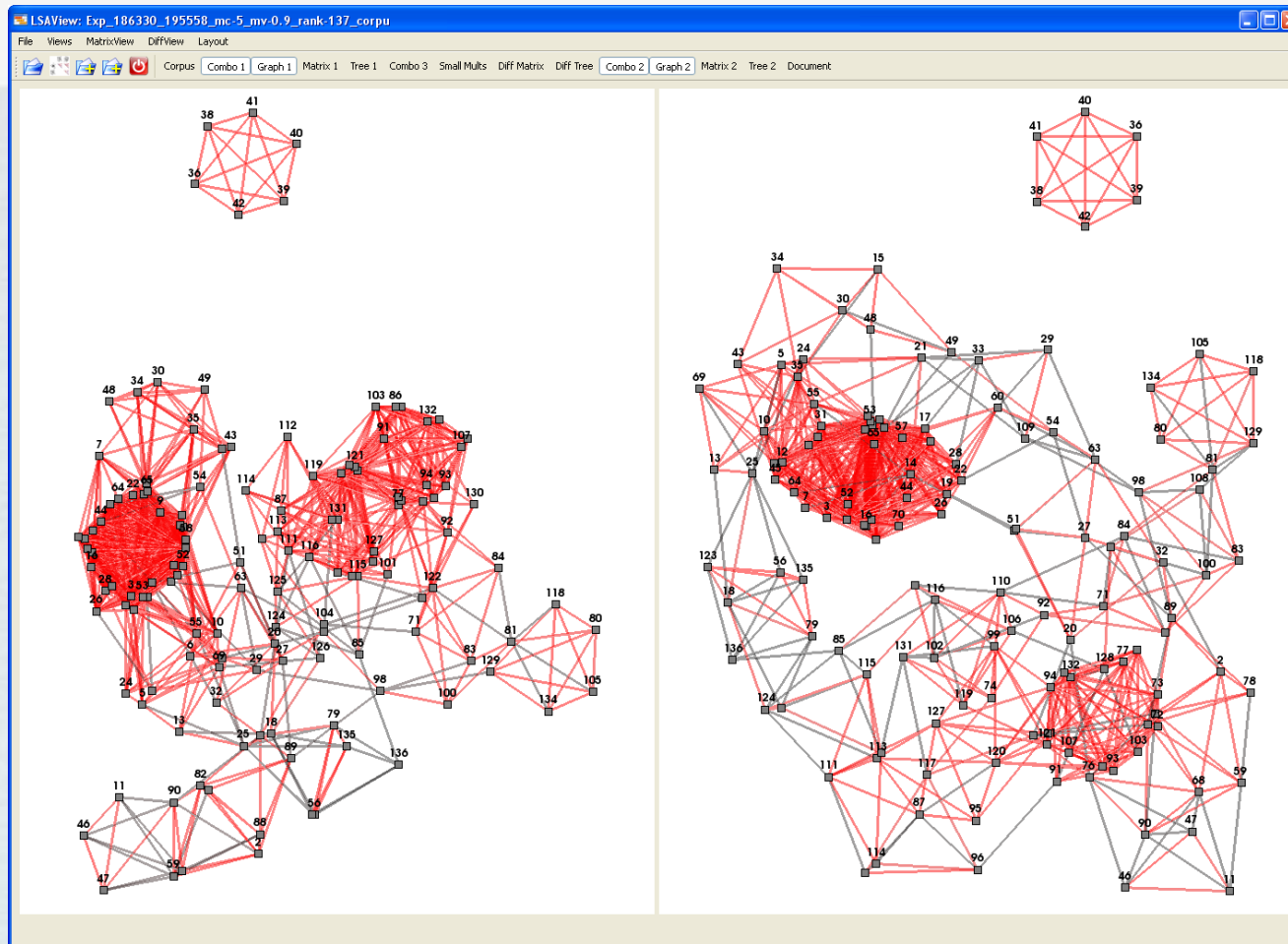


- Matrix views show edge weights
- $k = 100$  little difference in weights
- $k = 20$  good clustering



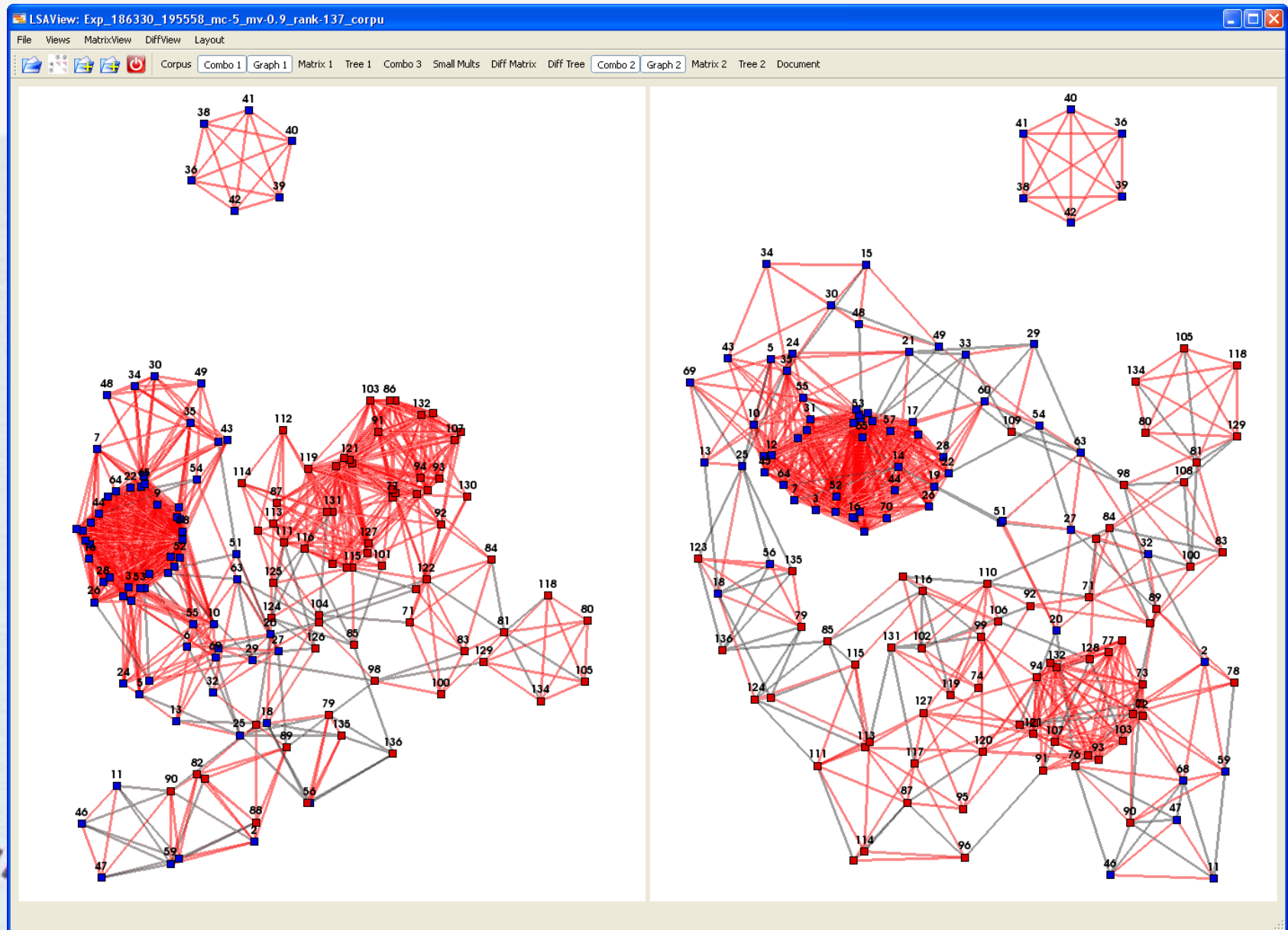


# TECHTC $k = 6$ , $\alpha = 1$ vs. $\alpha = -1$



- After further analysis, select  $k=6$
- Both  $\alpha$  have two distinct clusters
- Slightly stronger links in  $\alpha = -1$
- Both scalings perform well

# TECHTC True Cluster Assignments



# Conclusions

- Illustrated how LSAView used to understand LSA models
  - Seeding of other models (graph models)
  - Impact on document clustering task
- Key departure from previous work
  - Produces significantly different rank selection than automated approaches
  - Focuses on impact to text analysis tasks over variance



Work Funded by Laboratory Directed Research & Development (LDRD) program at Sandia National Laboratories

$$A_k = U_k \Sigma_k V_k^T$$

$$e_{ij}(k) = \frac{\langle v_k^i \Sigma_k, v_k^j \Sigma_k \rangle}{\|v_k^i \Sigma_k\|_2 \|v_k^j \Sigma_k\|_2}$$

$$e_{ij}(k, \alpha) = \frac{\langle v_k^i \Sigma_k^{\alpha/2}, v_k^j \Sigma_k^{\alpha/2} \rangle}{\|v_k^i \Sigma_k^{\alpha/2}\|_2 \|v_k^j \Sigma_k^{\alpha/2}\|_2}$$

$$\bar{e}_{ij}(k, \alpha, n) = \frac{1}{n+1} \sum_{r=k-n/2}^{k+n/2} e_{ij}(r, \alpha)$$

$$s_{ij}(k, \alpha, n) = \sqrt{\frac{1}{n} \sum_{r=k-n/2}^{k+n/2} (e_{ij}(r, \alpha) - \bar{e}_{ij}(k, \alpha, n))^2}$$



$$t_{ij}^{(1)} = \frac{\bar{e}_{ij}(k, \alpha, n) - e_{ij}(k, \alpha)}{s_{ij}(k, \alpha, n) / \sqrt{n+1}}$$

$$t_{ij}^{(2)} = \frac{\bar{e}_{ij}(k_1, \alpha, n_1) - \bar{e}_{ij}(k_2, \alpha, n_2)}{\sqrt{\frac{[s_{ij}(k_1, \alpha, n_1)]^2}{n_1} + \frac{[s_{ij}(k_2, \alpha, n_2)]^2}{n_2}}}$$